Why do I need four search engines?

Martin Holmes and Joseph Takeda

ABSTRACT

This presentation addresses the question of how to create digital editions and other online resources in forms that are likely to endure and remain functional over many decades. In particular, we focus on the dichotomy whereby the long-term robustness of a digital project, which is achieved by uncoupling it from transient server-side technologies and tools that require monitoring and maintenance, is undermined by the requirement to provide methods for users to search the collection. We present as a case-study the *Robert Graves Diary* project, which provides four separate search facilities using different approaches.

INDEX

1. Introduction

[1] Project Endings[1] is a collaboration of University of Victoria scholars, digital humanists and librarians whose aim is to address the progressive loss of digital scholarly resources due to failures in archiving, preservation, and documentation, and over-dependence on transient tools and technologies. The project is supported by a grant from the Social Sciences and Humanities Research Council of Canada (SSHRC).

[2] The project is working with a number of case-studies—digital edition projects already completed or nearing completion—and aiming to specify approaches, tools and technologies that can help researchers complete their projects and archive them in such a way that they have a strong chance of being available and functional for decades to come.

[3] In previous work (Arneil and Holmes 2017, Holmes 2017, Holmes and Takeda 2017), we have argued strongly that likelihood that a digital edition project will survive and be usable over the long term depends on the selection of a small core set of technologies (HTML5, CSS and JavaScript), and the avoidance of server-side technologies that will require maintenance or replacement over time. Our case-study projects (among them *Le Mariage sous L'Ancien Régime* and *The Robert Graves Diary*) are constructed entirely in this way, with no server-side dependencies at all.

2. The problem of search

[4] A digital edition consisting only of HTML, CSS and JavaScript can of course by rich and highly interactive. However, there is one important component of a website which generally requires some sort of server interaction: search. This is perhaps the most difficult challenge for the Endings project: how do we make a resource searchable without building in dependence on a server to host the index and respond to queries?

Figure 1. The first search interface of the Robert Graves Diary, which also serves as its home page.

5   Using the *Robert Graves Diary* project as a testbed, we have developed four distinct approaches, which will be described in this presentation:

1. *Bite the bullet and accept the server dependency.* We currently host the Graves project materials inside an eXist XML database, which allows us to provide rich faceted search functionality at the expense of a dependency that will inevitably be unsupported in the long term.

2. *Enlist Google's help.* We have built an additional Google Custom Search page into the site, allowing users to search in the interface which is probably most familiar to them. The obvious drawback here is that Google's terms, conditions and APIs change frequently, so we must expect this service to fail at some point when there is no active maintainer of the project.

3. *Enlist the help of our Library.* The long-term preservation of our project will ultimately be in the hands of the University Library, who run their own Solr server for searching their collections. As part of the project build, we now create JSON index files for Solr to ingest; we can then provide a search page which queries this index.

4. *Provide a standalone search.* For digital editions which are not too large, it is possible to create a JavaScript-only index, including stemming and relevance scoring, which is remarkably fast and requires no server support at all. This is the ultimate fallback when all else fails.

6   On the face of it, this level of redundancy may appear ridiculous, but in fact it provides a level of flexibility which we believe is essential for the survival of projects with no ongoing maintenance. In the best-case scenario, four different methods of searching the collection are available to the user, each with their own strengths and weaknesses. In the worst case, where the collection survives only as a simple collection of files on a drive somewhere, the standalone search will still work, while the other non-functional search interfaces provide evidence of the aspects of the collection thought to be crucial search facets.

BIBLIOGRAPHY

Arneil, Stewart and Martin Holmes. 2017. "Archiving form and function: preserving a 2003 digital project." DPASSH Conference 2017: Digital Preservation for Social Sciences and Humanities, Brighton, UK, 14th June 2018.

Holmes, Martin. 2017. "Selecting Technologies for Long-Term Survival." SHARP Conference 2017: Technologies of the Book, Victoria, BC, Canada, 10th June 2017. https://github.com/projectEndings/ Endings/raw/master/presentations/SHARP_2017/mdh_sharp_2017.pdf.

Holmes, Martin and Joseph Takeda. 2017. "Beyond Validation: Using Programmed Diagnostics to Learn About, Monitor, and Successfully Complete Your DH Project." Digital Humanities 2017 Conference, Montreal, Canada, 1th August 2017. https://dh2017.adho.org/abstracts/140/140.pdf.

NOTES

1   https://github.com/projectEndings,   https://onlineacademiccommunity.uvic.ca/ endingsproject/

AUTHORS

MARTIN HOLMES

University of Victoria

JOSEPH TAKEDA

University of British Columbia